Yin *et al.*
European Journal of Medical Research (2024) 29:540
https://doi.org/10.1186/s40001-024-02115-9

European Journal of Medical Research

RESEARCH

Open Access

Improving methylmalonic acidemia (MMA) screening and MMA genotype prediction using random forest classifier in two Chinese populations

Zhe Yin[1†], Chuan Zhang[1,2,3†], Rui Dong[4†], Xinyuan Zhang[1], Yingnan Song[1,3], Shengju Hao[2], Zhongtao Gai[4], Bingbo Zhou[2], Ling Hui[2], Shifan Wang[2], Huiqin Xue[5], Zongfu Cao[1,3*], Yi Liu[4*] and Xu Ma[1,3*]

Abstract

Background Methylmalonic acidemia (MMA) is one of the most common hereditary organic acid metabolism disorders that endangers the lives and health of infants and children. Early detection and intervention before the appearance of a newborn's clinical symptoms can control disease progression and prevent or mitigate its serious consequences.

Methods 42,004 newborns from two Chinese populations were included in the study. The small molecular metabolite analytes were detected from the dried blood spot (DBS) samples by MS/MS. Genetic analysis of 68 Chinese MMA cases were performed by whole-exome sequencing and Sanger sequencing. Random forest classifiers (RFC) were constructed to improve the MMA screening performance and genotype prediction in two Chinese populations. Meanwhile, other six machine learning models were trained to separate MMA patients from normal newborns. Model performance was assessed using accuracy, sensitivity, specificity, false positive rate (FPR), and positive predictive value (PPV) and the area under the receiver operating characteristic curve (AUC).

Results In the total 42,004 newborn samples, 68 MMA cases were identified by genetic analysis, 42 cases of which were caused by variants in *MMACHC*, 24 cases by variants in *MMUT*, and two cases by variants in *MMAA*. Three novel variants including c.449T>G (p.I150R) of *MMACHC*, c.1151C>T (p.S384F) and c.1091_1108delins (p.Y364Sfs*4) in *MMUT* were identified in the MMA patients. RFC for newborn screening of MMA performed best as compared to several other classification models based on machine learning with 100% sensitivity, low FPR, excellent PPV and AUC. In addition, the subdivision RFC for MMA genotype prediction was constructed with superior performance.

Conclusions It can be seen that RFC is extremely helpful for detection and genotype prediction in the newborn MMA screening. In addition, our findings extend the variant spectrum of genes related to MMA.

[†]Zhe Yin, Chuan Zhang and Rui Dong have contributed equally to this work.

*Correspondence:
Zongfu Cao
zongfu_cao@163.com
Yi Liu
y_liu99@sina.com
Xu Ma
maxubioinfo@163.com
Full list of author information is available at the end of the article

Keywords Newborn screening, Methylmalonic acidemia (MMA), Genotype, Random forest classifier, Machine learning

Introduction

Methylmalonic acidemia is a group of inborn errors of metabolism causing multisystem disease, the affected patients may have developmental, metabolic, hematological, ophthalmological, neurological, and dermatological clinically abnormal findings [1]. MMA is a genetically heterogeneous disease, isolated MMA can be caused by mutations in *MMUT* (OMIM# 609058), *MMAA* (OMIM# 607481), *MMAB* (OMIM# 607568), *MMADHC* (OMIM# 611935). MMA with homocystinuria can be caused by mutations in *MMACHC* (OMIM# 609831), *MMADHC* (OMIM# 611935), *LMBRD1* (OMIM# 612625), *ABCD4* (OMIM# 603214), *HCFC1* (OMIM# 300019). Except for the *HCFC1* gene, which is an X-linked recessive inheritance, other MMA disease-causing genes are autosomal recessive. Genotyping children with MMA is of great significance.

The incidence of MMA is between 1/48,000 and 1/250,000, which varies from country to country [2, 3]. In China, the most common form of organic aciduria is MMA. An estimated 1 in 26,000 infants are born with MMA in Shanghai and Beijing, China [4], but the incidence rate of *MMACHC* genotype is up to 1/3920 in Shandong, China [5]. Nearly, 30 genetic metabolic diseases including MMA [6] can be detected by tandem mass spectrometry (MS/MS), because MS/MS can simultaneously detect dozens of disorders of amino acid, organic acid, and fatty acid metabolism with one blood sample. Early detection and intervention before the appearance of newborn's clinical symptoms can control disease progression, prevent and mitigate the serious consequences.

Machine learning in medical examination and diagnosis can provide new ideas for improving diagnosis accuracy, guiding treatment decision-making, and improving patient management [7–10]. Furthermore, some studies have already used these machine learning techniques to improve clinical prediction of neonatal metabolic disorder-related diseases and have reported an improved classification accuracy [11–13]. However, such machine-learning models have not been widely used in pediatric clinical practice. The most pressing problem to be solved is to establish and fine-tune clinically acceptable classification models with 100% sensitivity, low FPR, and excellent PPV in different large populations with different genetic metabolic diseases. As a supervised machine learning algorithm, random forest (RF) is promising for applications in newborn screening (NBS) analysis. As an illustration, a RF classifier (RFC) was employed to develop Newborn Screening (NBS) models for several metabolic disorders, including ornithine transcarboxylase deficiency (OTCD), glutaric acidemia type 1 (GA-1), Methylmalonic acidemia (MMA) and very long-chain acyl-CoA dehydrogenase deficiency (VLCADD) [14]. RFC has been successfully used to improve phenylketonuria (PKU) screening performance with excellent sensitivity, false positive rate (FPR), and positive predictive value (PPV) on 41 MS/MS analytes in two Chinese populations [15].

Although the RFC model for MMA screening has been reported, the performances of sensitivity and other evaluation indicators need further improvement. In this study, we developed an RFC for newborn screening of MMA with 100% sensitivity, low FPR, excellent AUC, and PPV in two Chinese populations. Besides, we established an RFC for MMA genotype prediction using RFC with satisfied performance.

Materials and methods

Subject selection

A total of 42,624 newborns were collected, with 23,143 samples obtained from Gansu Provincial Maternity and Child-care Hospital in southeastern China between 2017 and 2020, and 19,481 samples collected from Children's Hospital Affiliated with Shandong University in eastern China from 2016 to 2021. All samples carrying information on other genetic metabolic disorders or treatment were excluded, comprising 51 MMA patients (46 from Shandong, 5 from Gansu) with treatment details and 569 patients (124 from Shandong, 445 from Gansu) diagnosed with various other metabolic disorders. The 51 excluded MMA patients included cases where MMA developed during the neonatal period and treatment began before screening. The 569 individuals excluded due to other metabolic disorders had received a definitive diagnosis through genetic testing. These disorders included conditions such as phenylketonuria, maple syrup urine disease, and homocystinuria, among others. The control group consisted of non-affected individuals who were matched for age, sex, and geographical location. These individuals had no known metabolic disorders and did not present with any screening abnormalities indicative of MMA or other metabolic conditions. Each MMA patient has a definite pathogenic mutation confirmed

by Sanger or Next-generation sequencing. This study protocol was approved by the Ethics Committee of the National Research Institute for Family Planning (Beijing, China). The personal information for all newborn samples was deleted to safeguard each person's privacy.

Metabolic analytes of the newborn screening

The small molecular metabolite analytes were detected from the dried blood spot (DBS) samples by MS/MS. There are 45 features in total including 10 amino acids metabolic analytes, including Alanine (Ala), Glycine (Gly), Proline (Pro), Valine (Val), Methionine (Met), Phenylalanine (Phe), Tyrosine (Tyr), Citrulline (Cit), Ornithine (Orn), Arginine (Arg); 31 fatty acids metabolic analytes, including Free carnitine (C0), Acetyl-carnitine (C2), Propionyl-carnitine (C3), Dicarboxybutyl-carni tine (C5OH+C4DC), Butyryl-carnitine (C4), Dicarboxypropyl-carnitine (C4OH+C3DC), Octenoic-carnitine (C5), Isohexenoyl-carnitine (C5:1), Pentadecanedioyl-carnitine (C5DC+C6OH), Hexanoyl-carnitine (C6), Octenoic-carnitine (C8:1), Octanoyl-carnitine (C8), Decadienoic-carnitine (C10:2), Decenoic-carnitine (C10:1), Decanoyl-carnitine (C10), Dodecenoyl-carnitine (C12:1), Dodecanoyl-carnitine (C12), Tetradecadienoic-carnitine (C14:2), Tetradecenoic-carnitine (C14:1), Tetradecanoyl-carnitine (C14), Hydroxytetradecanoyl-carnitine (C14OH), Hexadecenoic-carnitine (C6:1), Hexadecanoyl-carnitine (C16), Hydroxyhexadecenoic-carnitine (C16:1OH), Hydroxyhexadecanoyl-carnitine (C16OH), Octadecadienoic-carnitine (C18:2), Octadecenoic-carnitine (C18:1), Octadecanoyl-carnitine (C18), Hydroxyoctadecenoic-carnitine (C18:1OH), Hexanedioyl-carnitine (C6DC), Hydroxyoctadecanoyl-carnitine (C18OH); and the ratios of 4 fatty acids, which are Propionyl-carnitine/Free carnitine (C3/C0), Propionyl-carnitine/Acetyl-carnitine(C3/C2), Propionyl-carnitine/Hexadecanoyl-carnitine (C3/C16), and Phenylalanine/Tyrosine (Phe/Tyr). Phe/Tyr is a common indicator in NBS, and C3/C2, C3/C0, and C3/C16 are closely related to MMA [16]. Descriptive statistical information for each of the 45 features utilized in this study is provided in Additional file 1: Table S1.

Genetic analysis of MMA patients

Genomic DNA preparation

A total of 2–3 ml of blood samples were collected from the probands and their parents. Genomic DNA was extracted using the Tiangen Biotech DNA extraction kit (Beijing, China).

Whole-exome sequencing (WES)

Whole exome sequencing was performed using an Agilent SureSelect Human All Exon V6 Kit (Agilent Technologies Inc., USA) on an Illumina NovaSeq 6000 platform (Illumina Inc., CA, USA). This capture sequencing provides approximately 99% coverage of the target sequence, with an average depth > 20× coverage of 99%. Variants were described according to the nomenclature recommended by the Human Genome Variation Society (www.hgvs.org/). Variant frequencies were searched for in the GnomAD (http://gnomad.broadinstitute.org/), Exome Sequencing Project (ESP, http://evs.gs.washington.edu) and SNP (dbSNP) (http://www.ncbi.nlm.nih.gov/projects/snp) databases. Candidate variants were confirmed in the parents of each family by Sanger sequencing. Variants were checked in the Human Gene Variant Database (www.hgmd.cf.ac.uk) and ClinVar database (www.ncbi.nlm.nih.gov/clinvar/). InterVar (http://wintervar.wglab.org/) software was used to evaluate the pathogenicity of all variants according to the standards and guidelines of the American College of Medical Genetics and Genomics (ACMG) [17].

Data analysis

The results obtained from WES were compared with the reference genome (GRCh37/hg19), and the detected high-quality variants were annotated with variant information according to the PGenomics platform (https://pgenomics.cn), a national shared service platform for human genetic resources. What's more, PGenomics platform can combine information on the clinical phenotype, inheritance pattern, family codisjunction, and pathogenicity of the various point of the affected children to screen the detected candidate variation point and score them comprehensively by bioinformatics software; the higher the score, the higher the correlation. All the candidate causing variants were checked manually.

Sanger sequencing

Candidate variants were confirmed in the parents in each family by Sanger sequencing. PCR products were bi-directionally sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, USA) on an ABI 3500DX Genetic Analyzer (Applied Biosystems) after purification on 2% agarose gels.

Metabolic data sets and data processing

The entire study population was stratified into two distinct categories: individuals diagnosed with MMA and healthy controls devoid of the condition. Within the assemblage of MMA subjects, three primary genotypes

were identified: *MMUT*, *MMACHC*, and *MMAA*. Upon through completion of the specified preprocessing procedures, the data set originating from Shandong province, encompassing a cohort of 41 MMA patients (*MMACHC*, 26; *MMUT*, 14; and *MMAA*, 1) and 19,270 normal newborns, underwent a rigorous randomization process, with a 7:3 allocation ratio, to establish both training and validation sets in MMA screening. As a result, the training data set comprised 13,486 normal newborns and 31 MMA patients, while the validation set consisted of 5784 normal newborns juxtaposed against 10 MMA patients. Separately, an independent testing set was derived from Gansu Province, incorporating a total of 22,693 samples, including 27 MMA patients (*MMACHC*, 16; *MMUT*, 10; and *MMAA*, 1) and 22,666 normal newborns. Due to the scarcity of MMAA genotype data, the two patients with *MMAA* genotypes were excluded, and only 66 patients (Shandong, 40; Gansu, 26) carrying *MMACHC* and *MMUT* genotypes were selected for constructing the subdivision model of MMA genotype prediction. Following the training of various models on the training set, the performance of each model is measured and judged using the validation set. Therefore, the validation set can be used for model selection. The testing set is only used once after training to evaluate the generalization of the model. Figure 1 depicts the general process of data analysis in MMA screening and genotype prediction, offering a comprehensive overview of the sequential stages involved, from initial data acquisition to the final evaluation stage.

Models training

Seven machine learning models were trained to separate MMA patients from normal newborns, including Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Stochastic Gradient Descent (SGD), Multi-Layer Perceptron (MLP), and RFC. The 45 feature variables were input into the machine learning models as continuous variables. For all machine learning models, these 45 features were directly input into the models without undergoing any further transformation. All models were optimal models after parameter adjusting and were computed using Scikit-learn 1.0.1 in Python 3.8.5.

Random forest classifier

RFC is a combined classifier algorithm proposed by Breiman in 2001 [18]. It is a supervised learning algorithm and an ensemble learning algorithm based on the decision tree. The RFC constructs an ensemble of k trees, each trained on a bootstrapped data set subset via Bagging. At each node split, a random subset of features is considered to enhance diversity. Trees are grown to maximize node purity without pruning. Prediction involves aggregating outputs from all k trees for improved decision-making.

To achieve the optimal RFC model, the number of trees in the forest, a maximum depth, a minimum split size, and a minimum leaf sample size were fine-tuned by the Python library's "Grid Search" in this study. Due to the imbalance between MMA and non-MMA samples, we generated category weights with high weights for small number of samples and low weights for large number of samples. The ideal criterion for clinical practice in MMA screening is to simultaneously detect all MMA patients with excellent PPV. Each decision tree determines the MMA or non-MMA status of its own sample when one is added to the MMA screening model. By combining the disease status of each decision tree and employing a straightforward voting method with the minority following the majority, the model decides whether the sample is an MMA patient.

Similarly, each decision tree determines the *MMACHC* or *MMUT* status of its own sample when one is added

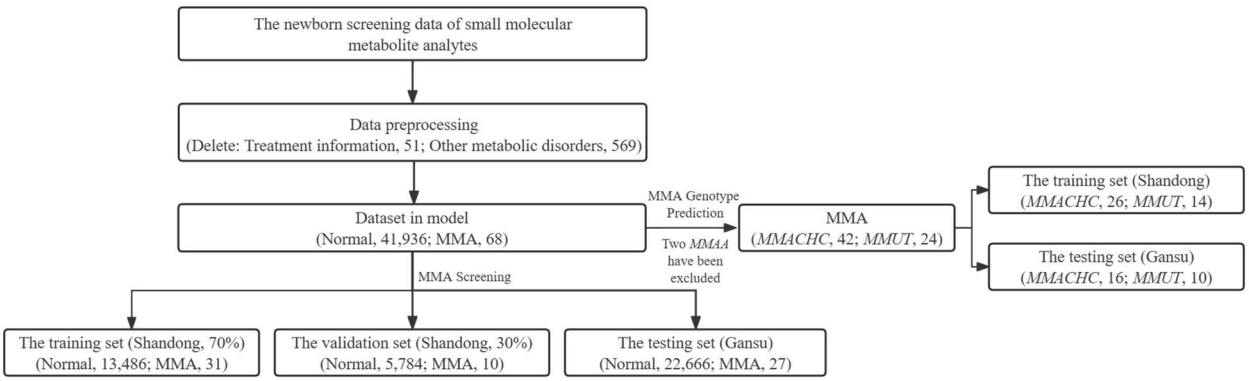

Fig. 1 General process for building an MMA screening and MMA genotype prediction model on MS/MS data

to the MMA genotype prediction model. By combining the *MMACHC* or *MMUT* status of each decision tree and employing a straightforward voting method with the minority following the majority, the model decides whether the genotype of the MMA patient is *MMACHC* or *MMUT*.

Feature importance

Gini impurity is the likelihood of incorrectly classifying a randomly chosen element in a data set based on its class distribution. The feature importance in RF represents the total reduction of Gini impurity on all nodes split based on the feature. The lower the Gini impurity, the higher the purity, the higher the order of the collection, and the better the classification effect.

Model performance

This work uses RFC to solve a binary classification problem. To see the accurate and incorrect class of each MMA status of the sample, the confusion matrix is employed (Table 1).

We employed an imbalanced data set for MMA screening. Since there are significantly more non-patient records than there are MMA patient records, accuracy cannot be the only metric. Then we assessed the performance of the classification using accuracy, sensitivity, specificity, false positive rate (FPR), and positive predictive value (PPV), as shown in Eq. (1). In addition, the area under the receiver operating characteristic curve (AUC) was utilized to evaluate the performance of the model, and the area under the receiver operating characteristic (ROC) curve is called AUC:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}},$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \quad (1)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

The Pearson chi-square test is used to test whether two categorical variables are independent of each other. It is a hypothesis-testing technique based on the chi-square distribution.

Table 1 Confusion matrix in MMA screening

Hypothesized class	True class	
	MMA	Non-MMA
MMA	True positives (TP)	False positives (FP)
Non-MMA	False negatives (FN)	True negatives (TN)

Results

Genetic analysis of MMA patients

In the total 42,004 newborn samples, we identified 68 MMA cases by genetics analysis, 42 cases were caused by variants in *MMACHC*, 24 cases were caused by variants in *MMUT*, and two cases were caused by variants in *MMAA*. In the *MMACHC* gene, a total of 20 different variants were detected, among which c.609G>A and c.656_658delAGA had the highest frequencies, 50% and 15.9% respectively. In the *MMUT* gene, a total of 30 different variants were detected, among which c.729_730insTT had the highest frequencies (8.3%), followed by c.626dupC (6.3%), c.1106G>A (6.3%), and c.278G>A (6.3%). In two patients with MMA caused by variants in the *MMAA*, a homozygous variant c.1076G>A/c.1076G>A was detected in one case, and a compound heterozygous variant c.988C>T/c.734-7A>G was detected in another case (Additional file 2: Table S2).

Among the variants detected in this study, the variant c.449T>G of MMACHC, c.1151C>T and c.1091_1108delins of *MMUT* were novel that not have been reported. According to the American College of Medical Genetics guidelines, c.449T>G of MMACHC was categorized as "pathogenic", c.1151C>T of *MMUT* was categorized as "pathogenic" and c.1091_1108delins of *MMUT* was categorized as "pathogenic" (Table 2).

Model selection for the newborn screening

It can be seen that DT, LR, and RFC in the validation set meet the basic requirements of MMA screening,

Table 2 Pathogenicity analysis of novel variants

Gene	Nucleotide change	Amino acid change	Pathogenicity	Conservative	ACMG evidence
MMACHC	c.449T>G	p.I150R	Likely pathogenic	Yes	PM2, PM3, PP3, PP4
MMUT	c.1151C>T	p.S384F	Likely pathogenic	Yes	PM2, PM3, PP3, PP4
MMUT	c.1091_1108delins	p.Y364Sfs*4	Pathogenic	Yes	PVS1, PM2, PM3, PP3, PP4

that is, the sensitivity is 100%. Two classifiers, LR and RFC, have achieved 100% sensitivity in the training, validation, and testing set (Table 3). The PPV and AUC of the RFC are higher than those of other classifiers, and the values in the testing set are 33.75% and 99.92% respectively. It follows that the RFC is the most appropriate model for MMA screening.

The final RFC in MMA screening model used 121 trees in the forest, a maximum depth of 4, and a minimum leaf sample size of 46. In the testing set, the AUC of the ROC curve reaches 0.9992 (Fig. 2A). The confusion matrix illustrating the predictive performance of the MMA screening within the testing set is presented in Fig. 2B. These results demonstrate the validity of the RFC as a clinically accepted screening tool for MMA screening.

Table 3 Evaluation results of various classifiers for MMA and all bold numbers represent better performance

Data sets	Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	AUC (%)
Training set	RFC	99.42	100.00	99.42	28.44	99.93
	LR	99.36	100.00	99.36	23.91	99.75
	SVM	99.48	100.00	99.48	23.91	99.75
	DT	99.45	100.00	99.44	22.68	99.82
	SGD	99.48	95.46	99.48	23.08	99.73
	MLP	100.00	100.00	100.00	100.00	100.00
Validation set	RFC	99.48	100.00	99.48	25.00	99.75
	LR	99.40	100.00	99.39	35.19	99.62
	SVM	99.40	94.74	99.41	34.62	99.63
	DT	99.00	100.00	99.00	23.00	100.00
	SGD	99.43	84.21	99.48	34.78	99.69
	MLP	99.78	36.84	99.98	87.50	93.39
Testing set	RFC	**99.77**	**100.00**	**99.77**	**33.75**	**99.92**
	LR	99.52	100.00	99.52	19.71	99.88
	SVM	99.67	85.19	99.68	24.21	99.34
	DT	99.68	88.89	99.70	25.81	94.27
	SGD	99.72	96.30	99.73	29.55	99.86
	MLP	100.00	37.00	100.00	88.00	93.00

LR linear regression, *SVM* support vector machine, *DT* decision tree, *SGD* stochastic gradient descent, *MLP* multi-layer perceptron

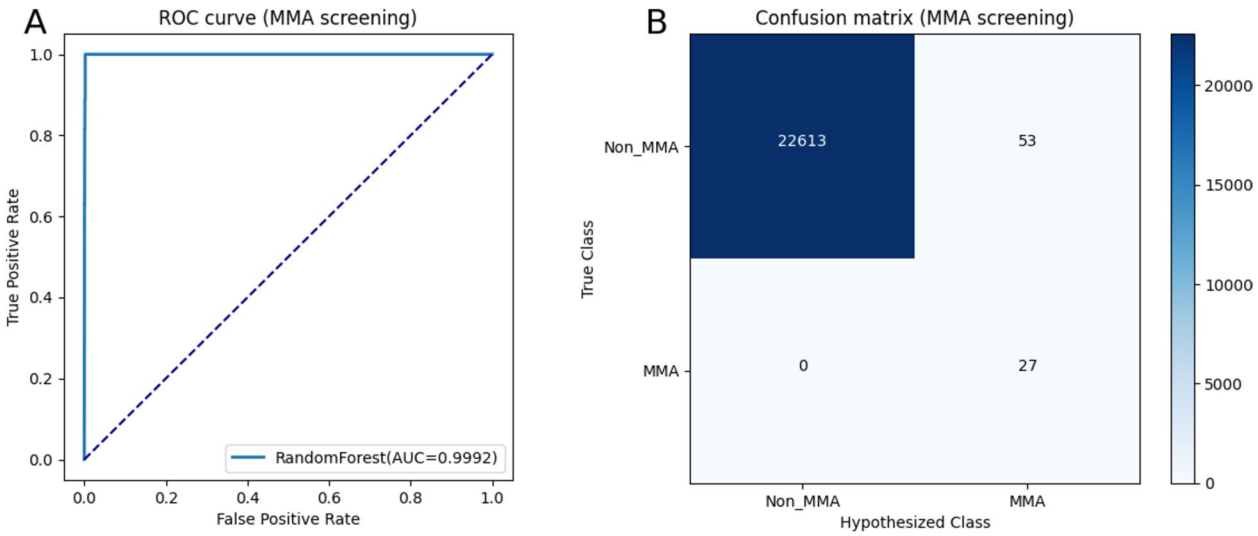

Fig. 2 **A** ROC curve using RFC analysis for MMA screening model in the testing set; **B** confusion matrix of the testing set for MMA screening using RFC

Figure 3 shows the importance ranking of all 45 features in MMA screening model. It can be known four of the top-ranked features, C3/C2, C3, C3/C0, and C3/C16, play critical roles in MMA screening.

The positive cases of MMA screening are determined by C3 acylcarnitine > 4.95 μmol/L or C3/C2 ratio > 0.27 in traditional clinical practice. PPV is 4.54% using the traditional screening method in the testing set. RFC reduced MMA false positive samples from 568 to 53 (9.33%) while PPV increased from 4.54 to 33.75% (Pearson's chi-squared test, $p = 0.0005$) in Table 4. Moreover, the RFC demonstrated exceptional discriminatory power with a specificity of 99.77% and a sensitivity of 100%, ensuring that no actual MMA cases were overlooked in the screening process. This compelling evidence underscores the RFC's profound

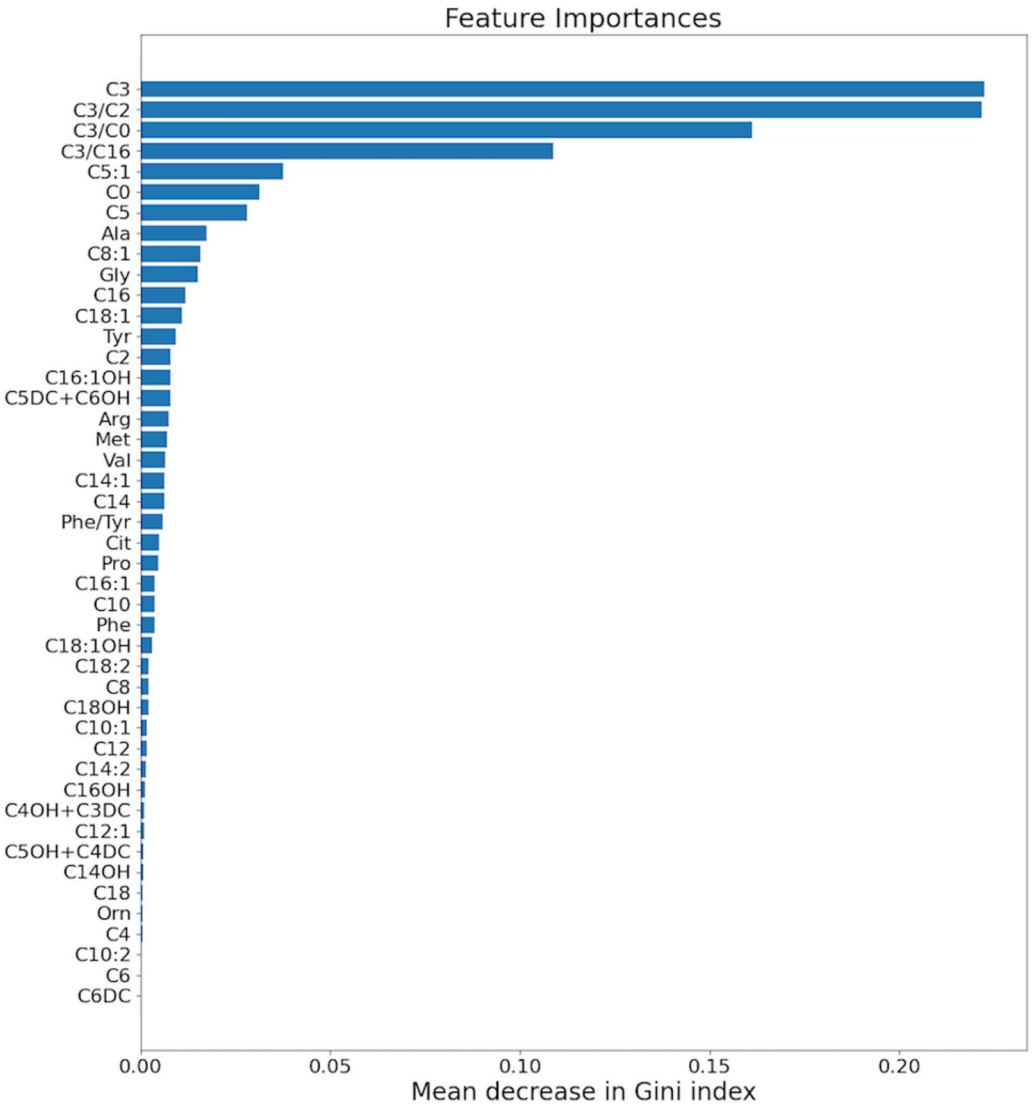

Fig. 3 Importance of 45 small molecular metabolite analytes in MMA screening using RFC

Table 4 Distribution and evaluation of RFC and traditional screening methods in independent testing set

Methods	TP	FP	TN	FN	Sensitivity (%)	Specificity (%)	PPV (%)
RFC	27	53	22,613	0	100.00	99.77	33.75
Traditional screening	27	568	22,098	0	100.00	97.49	4.54

impact on elevating the precision and effectiveness of MMA screening, affirming its capability to significantly bolster diagnostic accuracy.

RFC for MMA genotype prediction

The RFC for MMA genotype prediction used a forest of 25 trees, a maximum depth of 4, a minimum split size of 2, and a minimum leaf sample size of 1. The AUC of the ROC curve reaches 1.00 in the independent testing set (Fig. 4A). The confusion matrix of MMA genotype prediction in the independent testing set is depicted in Fig. 4B. It can be easily discovered the MMA genotype prediction model achieves excellent performance in the testing set. Of the 26 MMA patients, all patients are correctly predicted. These findings imply that the prediction of the MMA genotype can also benefit from the RFC. Figure 5 displays the importance ranking of all 45 features in the MMA genotype prediction model. It highlights that the top three ranked features-Met, C16:1OH, and Cit-play critical roles in accurately predicting MMA genotypes.

Discussion

In MMA screening, our RFC can detect all the MMA cases with lower the frequency of false positives. In all the training, validation, and testing sets, 100% sensitivity ensures that no MMA instances are missed. RFC performed best compared with other popular classification models including LR, SVM, DT, SGD, MLP, and RFC. In contrast to previous classification models, RFC demonstrated definite advantages. In the testing set, PPV greatly outperformed the conventional medical approach. The sensitivity needs to be 100% in the clinical situation to ensure that all MMA patients can be identified. This rule states that while LR, SVM, DT, and MLP algorithms perform well in the training data set, SVM, DT, and MLP struggle in the validation and testing set. Meanwhile, LR likewise performs exceptionally well, only slightly less than RFC.

RFC yields superior results in MMA screening and genotype prediction due to its utilization of ensemble learning. This method combines multiple decision trees to construct a more robust model, which helps reduce overfitting and improve generalization, thereby enhancing predictive accuracy. Additionally, the RFC provides a measure of feature importance that guides variable selection, enabling a focus on the most relevant predictive factors.

Gang Peng et al. [19] trained an RF analysis to enhance the prediction of true and false positives about MMA. Their model achieved performance in 96.12% sensitivity, 91% AUC, and 28.86% PPV in MMA screening. However, in our model, the performance of MMA screening was improved with 100% sensitivity, 99.92% AUC, and 33.75% PPV. Furthermore, our model completely eliminated the error rate, correctly classifying all 66 MMA patients without misclassification, unlike the 16% error rate observed in Peng et al.'s study where 15 out of 95 mut(0) (MMUT) patients were incorrectly classified as CblC (MMACHC), CblF (LMBRD1), or CblD (MMADHC).

A significant factor contributing to the excellence of our RFC models in terms of sensitivity, specificity, AUC, and PPV is the inclusion of two additional features, C3/C0 and C3/C16. These features, along with C3 and C3/

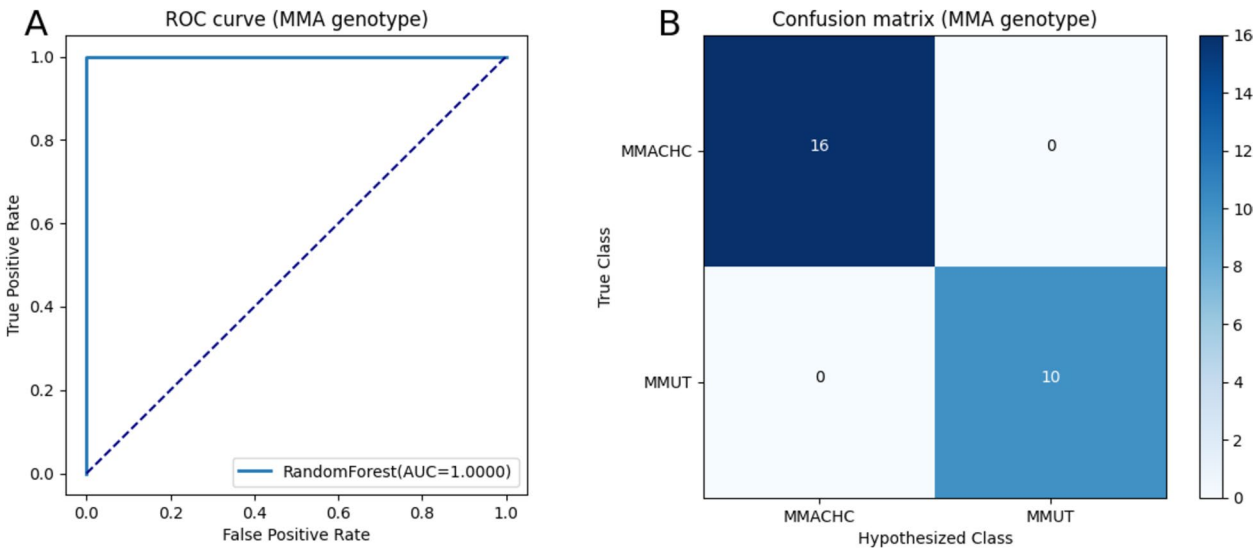

Fig. 4 **A** ROC curve using RFC analysis for MMA genotype prediction in dependent testing set; **B** confusion matrix of the independent testing set for MMA genotype prediction using RFC

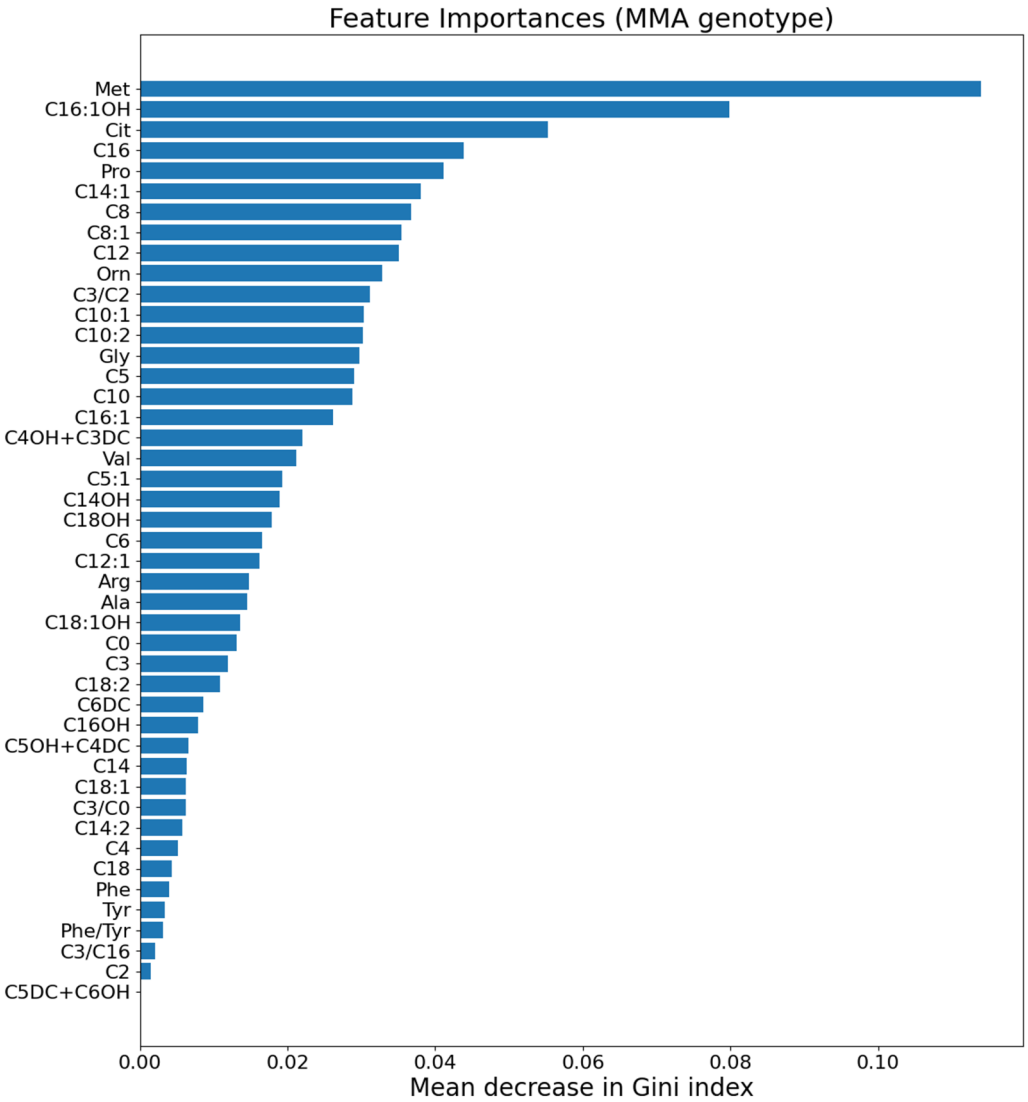

Fig. 5 Importance of 45 small molecular metabolite analytes in MMA genotype prediction using RFC

C2, were found to be highly important, as indicated by their lower Gini impurity. The absence of C3/C0 and C3/C16 in the RFC model could potentially lead to missed diagnoses of MMA patients and genotype misclassifications, as observed in [19]. According to the ranking of feature importance and excellent RFC performance in this study, C3/C0 and C3/C16 showed an essential function in MMA screening model. Our observations indicate that the primary cause of false-positive results in newborn screening is the levels of the primary MMA screening indicators C3, C3/C0, C3/C2, and C3/C16 being 5 to 10 times higher, or even higher, compared to those in patients with correct negative identifications.

The MMA genotype prediction model was applied to distinguish between *MMACHC* and *MMUT* genotypes. This model identified Met, C16:1OH, and Cit as the three most significant metabolic features. Met, a crucial amino acid, is typically reduced in *MMACHC* mutations; C16:1OH, a marker of mitochondrial dysfunction, is elevated in *MMACHC* mutation carriers; and Cit, a key intermediate in the urea cycle, whose levels reflect adjustments in energy metabolism related to the TCA cycle, is evident in both genotypes. These metabolites serve as effective diagnostic and monitoring biomarkers due to their significant changes in the presence of *MMACHC* and *MMUT* mutations.

In the classification problem of practical applications, the proportion of samples with different labels is likely to be unbalanced for the data set. If the algorithm training is directly used for classification, the training effect may be relatively poor. By adding the balanced category weight, the category weight will be inversely proportional to their frequency in the data, which can effectively solve the problem of sample imbalance. Due to the considerable disparity in data volume between positive and negative samples, we defined class weights for significantly imbalanced data in this study. Class weights are employed in the tree induction process to weigh Gini impurity in order to split [20]. This strategy is crucial as it enhances the capability of the MMA screening model, enabling it to maintain high accuracy in predicting the majority class while significantly improving the detection sensitivity for the minority class (MMA positive cases). By adopting this approach, the model treats all classes more equitably and, based on a limited number of rare disease samples, optimizes its ability to identify and distinguish rare cases. Consequently, it leads to an overall enhancement in diagnostic accuracy and practical utility.

Our research has some points worth discussing as well. First of all, the number of positive samples in the data set is insufficient due to the extremely low incidence of MMA. To validate the model, it is necessary to balance the number of negative and positive samples in the data set. Then, we recognized that the number of MMA patients included in our study was relatively small, so the RFC of MMA genotype prediction should be validated in more MMA patients. At the moment, the MMA genotype in our study consists of only two types: *MMACHC* and *MMUT*. Some rare disease-causing genes were not found in the two regions in this study, such as *HCFC1* and *ABCD4*. This might because in China, the MMA patients that caused by *HCFC1* less than 10 cases [21–23], and no MMA patients has been reported caused by *ABCD4*.

In the future, more MMA genotypes, including *MMACHC* and *MMUT*, will be collected in order to predict more MMA genotypes. At last but not least, a variety of factors can cause abnormalities in screening indicators. When applying tandem mass spectrometry technology to disease screening, an appropriate reference value range should be selected according to the different gestational ages and birth weights of newborns, blood collection time, and season. The interpretation of the measured value is frequently ambiguous in low birth weight and premature neonates, and there is currently no unambiguous reference value, which is bound to affect the predicted outcomes. Thus, it is a pressing issue to figure out how to reasonably reduce erroneous results and enhance screening effectiveness.

In addition, genotyping children with MMA is of great significance. After the genotyping is confirmed, it can accurately guide follow-up treatment and provide prenatal diagnosis and genetic counseling. We identified one novel variant in *MMACHC* and two novel variants in *MMUT*, which enlarged the variant spectrum of MMA-related genes. These findings help inform the genetic diagnosis of MMA and add to the theoretical basis for the prevention of MMA.

In the screening for MMA, differentiation from propionic acidemia (PA) is crucial, PA exhibits the same screening abnormality as MMA (e.g. C3, C3/C2). It is difficult to distinguish MMA from PA by MS/MS alone, and GC/MS is needed to further distinguish them. However, our study did not perform GC/MS to distinguish them, and this study only focused on MMA, which is the limitation of this study. For participants with abnormal C3, C3/C0, C3/C2 and other test indicators, if they were genetically diagnosed as non-MMA patients, we excluded them from the case group and control group. Further analysis is needed in the future to improve the disease-specific genotype prediction.

In conclusion, machine learning-based RFC can improve MMA screening performance with 100% sensitivity, low FPR, and excellent PPV in two Chinese populations. Similarly, RFC can accurately predict MMA genotype with very good performance. In conclusion, RFC is promising for the clinical application of MMA screening and MMA genotype prediction.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40001-024-02115-9.

Supplementary Material 1.

Supplementary Material 2.

Author contributions

All authors contributed to the work that is being presented here. Z.Y., C.Z., R.D., and Z.C. created this research. C.Z., R.D., S.H., Y.L., Z.G., B.Z., L.H., S.W., and H.X. provided the data. Z.Y., R.D., Y.S., C.Z., and X.Z. handled the data. Z.Y., C.Z., X.Z., Y.S., Z.C., Y.L., and X.M. carried out the statistical modeling and data analysis. Z.Y., C.Z., R.D., Y.S., Z.C., and X.M. wrote and reviewed the manuscript. All writers read and approved the final manuscript.

Funding

This work was funded by the National Key Research and Development Program of China (2016YFC1000307), National Human Genetic Resource Sharing Service Platform (2005DKA21300), Gansu Provincial Science and Technology Plan Funding Project (22YF7FA094) and Lanzhou Science and Technology Plan Project (2021-1-182).

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
The study protocol was approved by the Ethics Committee of the National Research Institute for Family Planning (Beijing, China). Before newborn screening and the study, each patient's parent or guardian provided informed consent.

Competing interests
The authors declare no competing interests.

Author details
[1]National Human Genetic Resources Center, National Research Institute for Family Planning, Beijing, China. [2]Gansu Province Medical Genetics Center, Gansu Provincial Clinical Research Center for Birth Defects and Rare Diseases, Gansu Provincial Maternity and Child-Care Hospital, Lanzhou, China. [3]Graduate School of Peking Union Medical College, Beijing, China. [4]Pediatric Research Institute, Children's Hospital Affiliated to Shandong University, Jinan, China. [5]Department of Cytogenetic Laboratory, Shanxi Children's Hospital, Shanxi Women and Children Hospital, Affiliated Hospital of Shanxi Medical University, Taiyuan, China.

Received: 19 April 2024 Accepted: 16 October 2024
Published online: 10 November 2024

References

1. Zhang C, Wang X, Hao S, Zhang Q, Zheng L, Zhou B, Liu F, Feng X, Chen X, Ma P, Chen C, Cao Z, Ma X. Mutation analysis, treatment and prenatal diagnosis of Chinese cases of methylmalonic acidemia. Sci Rep. 2020;10(1):12509.
2. Weisfeld-Adams JD, Bender HA, Miley-Akerstedt A, Frempong T, Schrager NL, Patel K. Neurologic and neurodevelopmental phenotypes in young children with early-treated combined methylmalonic acidemia and homocystinuria, cobalamin C type. Mol Genet Metab. 2013;110(3):241–7.
3. Lindner M, Gramer G, Haege G, Fang-Hoffmann J, Schwab KO, Tacke U, Trefz FK, Mengel E, Wendel U, Leichsenring M, et al. Efficacy and outcome of expanded newborn screening for metabolic diseases—report of 10 years from South-West Germany. Orphanet J Rare Dis. 2011;8:44.
4. Tu W, Chen H, He J. Methylmalonic aciduria: newborn screening in mainland China? J Pediatr Endocrinol Metab. 2013;26(3–4):399–400.
5. Han B, Cao Z, Tian L, Zou H, Yang L, Zhu W, Liu Y. Clinical presentation, gene analysis and outcomes in young patients with early-treated combined methylmalonic acidemia and homocysteinemia (cblC type) in Shandong province. China Brain Dev. 2016;38(5):491–7.
6. Chace DH, Kalas TA, Naylor EW. Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns. Clin Chem. 2003;49(11):1797–817.
7. Ardila D, Kiraly A, Bharadwaj S. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med. 2019;25(6):954–61.
8. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, Jastrzębski S, Févry T, Katsnelson J, Kim E, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging. 2020;39(4):1184–94.
9. Li M, Fu X, Li D. Diabetes prediction based on XGBoost algorithm. IOP Conf Ser Mater Sci Eng. 2020;768: 072093.
10. Peter KC, Shen X, Wang G, Ho C, Leung C, Ng C, Choi K, Teoh JY. Enhancement of prostate cancer diagnosis by machine learning techniques: an algorithm development and validation study. Prostate Cancer Prostatic Dis. 2022;25(4):672–6.
11. Baumgartner C, Bohm C, Baumgartner D, Marini G, Weinberger K, Olgemöller B, Liebl B, Roscher AA. Supervised machine learning techniques for the classification of metabolic disorders in newborns. Bioinformatics. 2004;20(17):2985–96.
12. Chen W, Hsieh S, Hsu K, Chen H, Su X, Tseng Y, Chien Y, Hwu W, Lai F. Web-based newborn screening system for metabolic diseases: machine learning versus clinicians. J Med Internet Res. 2013;15(5): e98.
13. Zaunseder E, Mütze U, Garbade SF, Haupt S, Feyh P, Hoffmann GF, Heuveline V, Kölker S. Machine learning methods improve specificity in newborn screening for isovaleric aciduria. Metabolites. 2023;13(2):304.
14. Peng G, Tang Y, Cowan TM, Enns GM, Zhao H, Scharfe C. Reducing false-positive results in newborn screening using machine learning. Int J Neonatal Screen. 2020;6(1):16.
15. Song Y, Yin Z, Zhang C, Hao S, Li H, Wang S, Yang X, Li Q, Zhuang D, Zhang X, et al. Random forest classifier improving phenylketonuria screening performance in two Chinese populations. Front Mol Biosci. 2022;9: 986556.
16. Interlaboratory Quality Evaluation Committee of Neonatal Genetic and Metabolic Disease Screening, Clinical Test Center, Ministry of Health, Beijing Hospital, et al. Expert consensus on tandem mass spectrometry screening technology for neonatal diseases. Chin J Lab Med. 2019;42(2):89–97.
17. Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am J Hum Genet. 2017;100(2):267–80.
18. Breiman L. Random forests. Mach learn. 2001;45:5–32.
19. Peng G, Shen P, Gandotra N, Le A, Fung E, Jelliffe-Pawlowski L, Davis RW, Enns GM, Zhao H, Cowan TM, et al. Combining newborn metabolic and DNA analysis for second-tier testing of methylmalonic acidemia. Genet Med. 2019;21(4):896–903.
20. Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. 2004.
21. Wang F, Liang L, Ling S, Yu Y, Chen T, Xu F, Gong Z, Han L. Clinical characteristics and genotype analysis of five infants with cblX type of methylmalonic acidemia. J Zhejiang Univ (Med Sci). 2022;51(3):298–305.
22. Li D, Liu Y, Ding Y, Li X, Song J, Li M, Qin Y, Yang Y. A pedigree of a rare Cb1X type X-linked methylmalonic acidemia due to transcriptional co-regulator HCFC1 mutation. J Clin Pediatr. 2016;34(3):212–6.
23. Shen Y, Hu Z, Yang J, Yang R, Huang X. A case of methylmalonic acidemia and homocysteinemia cblX type with negative tandem mass spectrometry testing. J Zhejiang Univ (Med Sci). 2021;50(6):795–8.

Publisher's Note